An Internet-enabled Knowledge Discovery Process

An Internet-enabled Knowledge Discovery Process*

Alex G. BÜCHNER[α], Maurice D. MULVENNA[β], Sarab S. ANAND[χ], John G. HUGHES[α]

[α] Northern Ireland Knowledge Engineering Laboratory
[χ] School of Information and Software Engineering
Faculty of Informatics
University of Ulster
{ag.buchner, ss.anand, jg.hughes}@ulst.ac.uk

[β] MINE*it* Software Ltd.
Faculty of Informatics
University of Ulster
maurice@mineit.com

Abstract

A holistic approach, in the form of a process, is proposed in order to discover marketing intelligence from Internet data. The Internet-enabled knowledge discovery process contains the steps human resource identification, problem specification, data prospecting, domain knowledge elicitation, methodology identification, data pre-processing, pattern discovery, and knowledge post-processing. It also involves the three types of expertise required during a project, namely a web administrator, a marketing expert, and a data mining specialist. To show the validity and applicability of the proposed approach, the electronic commerce marketing scenarios of customer attraction, customer retention, cross-sales, and churn are tackled with the outlined process, respectively.

1. Introduction

Knowledge discovery of Internet data (also known as web mining), has been an area of recent cross-disciplinary research interest. For the discovery of patterns that are actionable in electronic commerce environments, web usage mining, that is the mining of server log files and related marketing information, has proved to be the most appropriate paradigm. Although many techniques have been proposed in the area of Internet data pre-processing (Cooley *et al*., 1999), data consolidation (Zaïane *et al*., 1998), and pattern discovery (Cooley *et al*., 1999, Büchner & Mulvenna, 1998), no process has been developed which covers the entire life-cycle of an online customer, the available operational and materialised data, as well as the incorporation of marketing knowledge.

The objective of this paper is to remedy this shortcoming and to propose an Internet-enabled knowledge discovery process. This not only covers various stages of a knowledge discovery exercise, but also involves experts which are involved in each project, as well as their specific knowledge.

---

* This research has partly been funded by the ESPRIT project Nº 26749.

The paper is outlined as follows. In Section 2, the Internet-enabled knowledge discovery process is proposed, before each step is described in detail. In Section 3, the applicability of the process is shown using the electronic commerce marketing scenarios customer attraction, customer retention, cross-sales, and churn, which are tackled with the suggested process, respectively. Related work is evaluated in Section 4, before, in Section 5, conclusions are drawn and further work is outlined.

2. The Knowledge Discovery Process

In order to facilitate all knowledge discovery steps, a web-enabled knowledge discovery process[1] has been developed (see Figure 1), which is an adoption of a generic process defined in earlier work (Anand & Büchner, 1998). There exists a vast amount of management problems, which initiate an electronic commerce-related knowledge discovery project. The sequence of steps which such a project should follow is outlined in the following sub-sections. The objective of each task is described in general first, before explicit electronic commerce scenarios are presented.

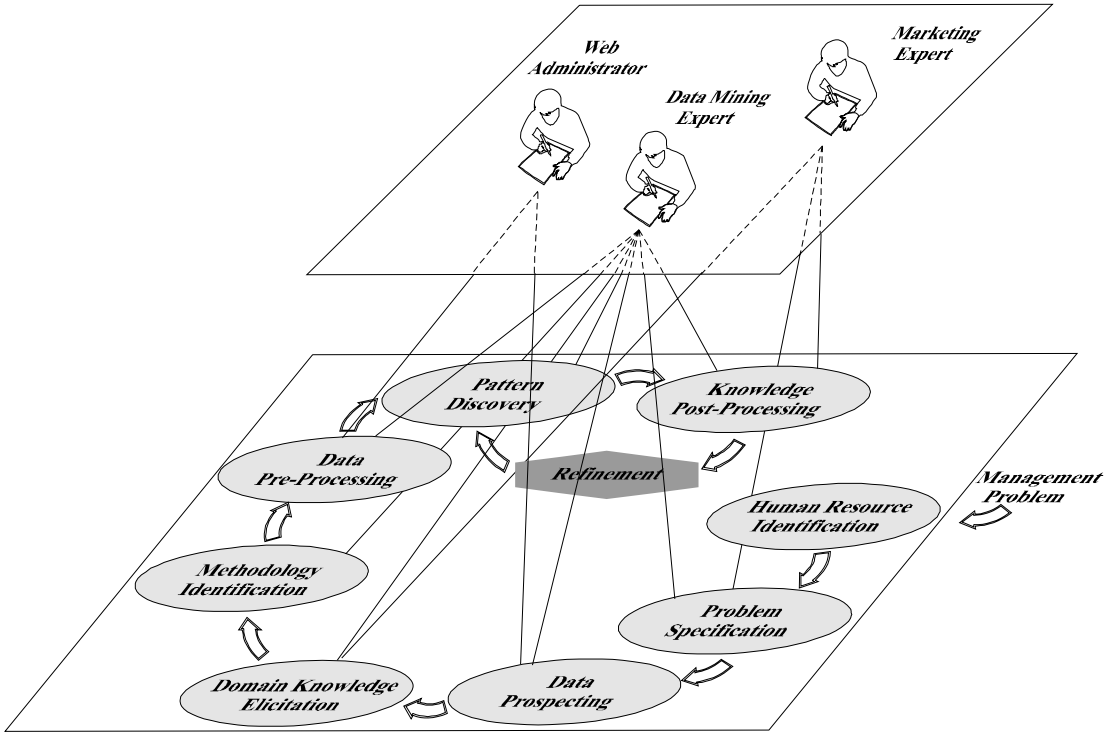

Figure 1: Internet-enabled Knowledge Discovery Process

[1] The terms web-enabled, Internet-enabled and electronic commerce-enabled are used interchangeably throughout this paper.

2.1 Human Resource Identification

After a problem has been identified at the management level of a virtual enterprise, human resource identification is the first stage of the knowledge discovery process, which requires domain, data and data mining expertise. The synergy of these human resources as early as possible within any project is imperative to its success.

In the case of web mining, the expertise involved is a web administrator (providing knowledge about the physical and logical arrangement of the retail site), a marketing expert (providing already known marketing knowledge and strategies, as well as verifying the discovered patterns), and a data mining specialist (driving the process from a technology-orientated point of view). In an electronic commerce environment it is not unusual to find the personnel involved at different physical locations.

2.2 Problem Specification

At the problem specification stage, a better understanding of the problem is developed by the human resources identified in the previous stage. The problem is decomposed into tasks and those tasks that can be solved using a knowledge discovery approach are identified. Each of these tasks is associated with a particular knowledge discovery goal. There are many goals which can be achieved through the application of knowledge discovery techniques, for instance, association discovery, classification, cluster analysis, sequential pattern discovery, temporal modelling, deviation detection, regression, characteristics discovery, or dependency modelling (for more detailed information refer to Anand & Büchner, 1998). Typical electronic commerce-specific goals are

- find the 10% most common page sequences browsed;
- discover the 20% most prospective virtual customers for a cross-sales exercise;
- unveil typical characteristics that distinguish a browser from a buyer during a campaign.

2.3 Data Prospecting

In order to perform electronic commerce-specific data prospecting, an online retailing model has been developed, which contains the processes on an electronic retailer's web site, as well as the back-end data storage components, which is depicted in Figure 2 (Mulvenna *et al*., 1998). The online retailing model is the permutation selected to research the efficacy of data mining. The primary reasons for selecting the online retailing model are behavioural (consumers exhibit a wide range of differing behavioural patterns), marketing-orientated (consumer and product information provides a rich source of useful data), and descriptive (products and services in online retail sites are varied). The three types of data are described in more detail in the following sub-sections.

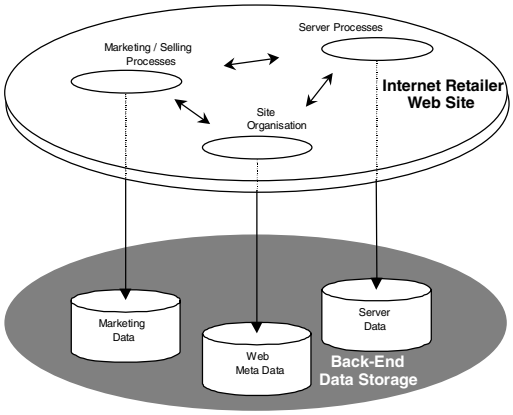

Figure 2: Online Retailing Model

2.3.1 Server Data

Server data is generated by the interactions between the persons browsing an individual site and the web server. The httpd process that runs on web servers provides a facility to log information on accesses to the server. That data can further be divided into log files and query data.

There are three types of log files, namely server logs, error logs, and cookie logs. Server logs are either stored in the Common Logfile Format or the more recent Extended Logfile Format. The available field information from the two files encompasses date and time request, remote host IP or DNS, user name, transferred byes, server name and IP, URI query and stem, http status, requested service name, time taken, transfer protocol version, user agent, cookie id, and referrer page. Additionally, the Extended Logfile Format supports directives which provide meta information about the log file, such as version, start and end date of session monitoring, as well as the fields which are being recorded.

Error logs store data of failed requests, such as missing links, authentication failures, or timeout problems. Apart from detecting erroneous links or server capacity problems — which, when satisfactorily corrected, can be seen as a compulsory form of customer satisfaction — the usage of error logs has so far proven rather limited for the discovery of actionable marketing intelligence.

Cookies are tokens generated by the web server and held by the clients. The information stored in a cookie log helps to ameliorate the transactionless state of web server interactions, enabling servers to track client access across their hosted web pages. The logged cookie data is customisable, which goes hand in hand with the structure and content of the marketing data (see Section 2.3.2).

A fourth data source that is typically generated on e-commerce sites is query data to a web server. For example, customers to an online store may search for products, or clients to a research database may search for publications. The logged query data must be linked to the access log through cookie data and / or registration information. There are currently no formal drafts for standards for handling query data, although new specification suggestions have reached draft

stage, for instance Resource Description Framework RDF. In order to make use of query data, it has to be grouped into logical, usually marketing-related clusters.

2.3.2 Marketing Data

Any organisation that uses the Internet to trade in services and products uses some form of information system to operate Internet retailing. Clearly some organisations use more sophisticated systems than others. The least common denominator information that is typically stored is about customers, products and transactions, each on different levels of detail. More sophisticated electronic traders keep also track of customer communication, distribution details, advertising information on their sites associated with products and / or services, sociographic information, and so forth.

2.3.3 Web Meta Data

The last source is data about the site itself, usually generated dynamically and automatically after a site update. Web meta data provides the topology of a site, which includes neighbour pages, leaf nodes and entry points. This information is usually implemented as site-specific index table, which represents a labelled directed graph. Meta data also provides information whether a page has been created statically or dynamically and whether user interaction is required or not.

In addition to the structure of a site, web meta data can also contain information of more semantic nature. Examples are arbitrary or ontology-based content information, usually represented through HTML meta tags or XML statements, the type of a page (root, navigational, content page, or a hybrid thereof), or page scores, which have been derived according to some pre-defined set of heuristics.

2.4 Domain Knowledge Elicitation

The main objective of domain knowledge elicitation and later incorporation at the pattern discovery stage is to constrain the learning algorithms search space and to reduce the number of patterns discovered. Marketing knowledge is a type of domain expertise, obtained internally or externally has usually been formulated by (human or artificial) marketing experts. It can be in the form of target-directed data collated from cross-fertilised sources, or the output of Internet mining activities carried out at an earlier stage.

More formalised domain knowledge (Anand *et al.*, 1995) is expressed in hierarchical generalisation trees or concept hierarchies (for example, the topological organisation of Internet domains), attribute relationship rules (constraints given by the marketing expert or discovered through data mining), attribute dependencies (the interdependency between a customer's URL and her/his postal address country), and environment-based constraints (the costs of recognising non-buyers among buyers, also known as having false negative errors). Internet-specific domain knwoledge includes networks, for instance, the topological organisation of a an e-tailer's web site and page chains, which specify certain browser paths of interest (Büchner *et al.*, 1999a).

2.5 Methodology Identification

The main task of the methodology identification stage is to find the best knowledge discovery methodology to solve the specified electronic commerce problem. Often a combination of paradigms is required to solve the problem at hand. Typical knowledge discovery methodologies are neural networks, rule induction, Bayesian belief networks, genetic algorithms, statistics, evidence theory, fuzzy and rough sets, and case-based reasoning. Details about those methodologies can be found in standard knowledge discovery literature such as Anand & Büchner (1998).

Knowledge discovery methodologies which have been applied in the context of direct marketing are rule induction (Ling & Li, 1998), neural networks (Bigus, 1996), statistics (Berry & Linoff, 1997), and genetic algorithms (Bhattacharyya, 1998).

2.6 Data Pre-processing

Section 2.3 has shown the highly arbitrary nature of available data sources in an electronic commerce environment. This section deals with the data preparation of that information, which has proven the most complex and time-consuming task in almost all web mining projects. The data pre-processing step consists of four sub-tasks, namely the resolution of schematic and semantic heterogeneities among the relevant data, a battery of data preparation activities, some of which are generic, but most are Internet- and electronic commerce-specific, the definition of a materialised view in form of a web log data hypercube, and the design of a snowflake schema.

2.6.1 Schematic and Semantic Heterogeneity Resolution

The types of data occurring in an electronic commerce scenario contain different types of heterogeneities, which are mainly stemming from the level of standardisation for each data source. Due to the fact that most server data sources are standardised, relatively little schematic heterogeneity occurs in these files. Less standardised and hence more unstructured is query data, since it is usually set up for individual, application- and domain-specific purposes. Similarly, type, content and structure of marketing information and meta data depend heavily on the electronic commerce domain, the topology of the site, the logical and physical interconnectivity with other sites (for instance, on shopping malls), and so forth. The type, level, and granularity of schematic heterogeneities among entities, as well as resolution thereof, is beyond the scope of this paper, and has been described in detail in Büchner & Mulvenna (1998).

After the creation of a conflict-free schema, various interpretations of available data sources are necessary. Typical examples in log files are date and time formats of logs being used in different countries, query information in different languages, or URIs which are interpreted differently in different contexts, for example, the suffix *.com* might represent commercial organisations in one situation, whereas in another context it might include private users as well (customers who are logged in through a service provider). Similar scenarios exist in marketing as well as web meta data, especially when data is distributed across different locations. In Büchner *et al*. (1999b) an

architecture has been described, which considers contextual information for the incorporation in a knowledge discovery process that has been applied in an e-commerce scenario.

2.6.2 Internet-specific Data Preparation Activities

In addition to generic data warehousing-like data preparation activities, some Internet- and electronic commerce-specific operations have to be carried out. First group consists of a set of techniques comprising cleansing, transforming, and aggregating. It is referred to in some of the standard literature (for instance, Berry & Linoff, 1997, Chaudhuri & Dayal, 1997) for more detailed information. For the purpose of this paper, it is focused on the second group.

Cooley *et al*. (1999) have created an interesting web mining architecture, which contains a range of sophisticated Internet data preparation tasks, each tailored towards a specific knowledge discovery goal (see also Section 4). From server logs, meta data files and some optional usage statistics, a user session file, a transaction file, the site topology, and page classifications are derived. Some activities are based on artificial intelligence techniques, for instance page classification is based on rule induction, whereas others use lookup tables, filtering (of multimedia information), and so on. Many of the outlined data preparation steps are inevitable for successful web usage mining. In order to handle all data sources as outlined in the previous section, essential operations were added in our process. Some of the techniques were expanded in order to handle extended log file entries and customer data were linked to a user session via cookies. Furthermore, some initial work has been carried out in considering ontology-based content information, expressed in XML.

2.6.3 Creating a Web Log Data Hypercube

After a rather informal description of data sources in electronic commerce scenarios and an arsenal of heterogeneity resolution and data preparation techniques, the design of the pre-processed and consolidated data is presented more formally. In order to create a materialised view which is used as repository for further analytical activities, an *n*-dimensional web log data cube is defined.

Definition 1. A web log data hypercube H represents an n-dimensional information space, such that $H = [D_1, D_2, D_3, \ldots]$, where each D represents a dimension of H. □

This hypercube represents an intentionally denormalised materialised view of the pre-processed data.

Definition 2. A dimension D represents a set of attributes such that $D = [a_{x_1}, a_{x_2}, a_{x_3}, \ldots]$. □

Some attributes may be derived fields in order to hold summarisation information, like resource hits, sales aggregation, or customer purchase sums.

Definition 3. Each cell is accessed through H[a_{2i2}, a_{2i2}, a_{3i3}, …]. The total number of cells is calculated as $|C| = \prod_{i=1}^{n} |H_i|$, where $|H_i|$ is the cardinality of H (number of attributes). □

This number leads to an exponential growth of the number of cells, many of which are naturally sparse, which is handled by internal compression mechanism. Also, efficient computation of web hypercubes has to be guaranteed. Harinarayan *et al*. (1996) covers both aspects, which are based on dynamic sparse matrixes; further details are beyond the scope of this paper. In order to simplify generalisation and aggregation operations as well as the modelling of dimensional hierarchies in snowflake schemas (see Section 2.6.4), each dimension is defined on a multi-level concept hierarchy (Büchner & Mulvenna, 1998).

Definition 4. A concept hierarchy T is an undirected, connected, acyclic graph which is defined as the tuple $T = (L, E)$, where $L = \{l_0, l_{1_1}, l_{1_2}, ..., l_{2_1}, l_{2_2}, ...\}$ and $E = \{e_1, e_2, e_3, ...\}$. Each l represents a value of a domain D of a hypercube H, such that the granularity $g(l_n) < g(l_{n-1})$, $n > 0$. Each e_k has the form $e_k = <l_i, l_j>$; $l_i, l_j \in L$, l_0 has indegree 0, $l_1...l_n$ have indegree 1. l_i is subconcept of l_j iff $l_i \subset l_j$; l_i is superconcept of l_j iff $l_j \subset l_i$. □

Thus, a hypercube can also be represented as an n-tuple $H=(T_1, T_2, T_3, \ldots)$. An example of a three-dimensional web log data cube as well as a drilled-down version with the dimensions customer, location and product is depicted in Figure 3.

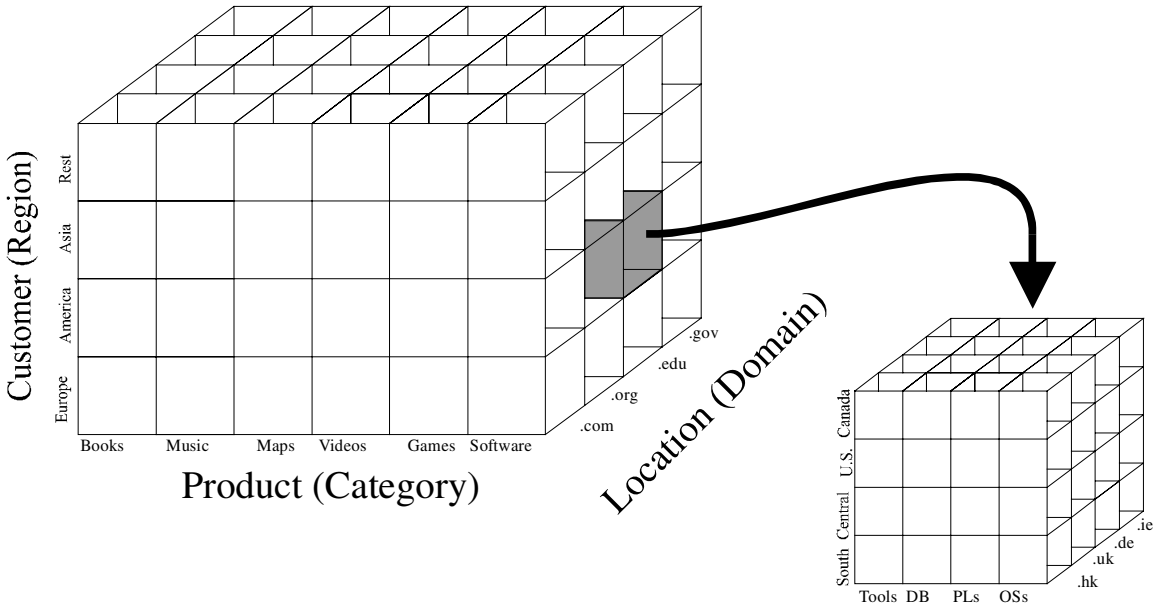

Figure 3: A three-dimensional Web Log Data Cube

2.6.4 Schematic Design

To construct a cube as depicted above and to support further analysis activities such as OLAP and data mining, a schema based on the relational calculus is modelled. Each dimension is represented as a relation, which is connected to a fact table. Fact tables act as connecting element in a data model representing keys and summarisation information. This star schema is sufficient for one given set of scenarios, which uses data input of the same granularity (Berson & Smith, 1997). For more advanced operations, as necessary in web mining, a snowflake schema is required, which supports multiple granularities. Snowflake schemas provide a refinement of star schemas where the dimensional hierarchy is explicitly represented by normalising dimension tables (Chaudhuri & Dayal, 1997).

Figure 4 depicts a web log snowflake schema that is centred around a fact table which is typical for analyses in e-commerce scenarios. It contains key fields (CustomerKey, ProductKey, LocationKey, DateKey, SessionKey) as well as some statistical summarisation information (Quantity, TotalPrice, ClickThroughRate).

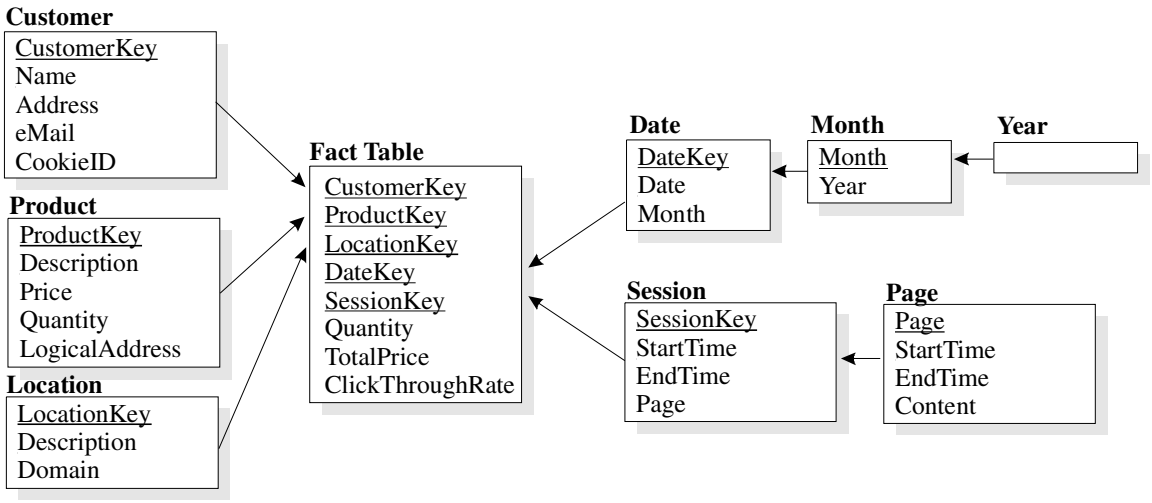

Figure 4: A Web Log Snowflake Schema

2.7 Pattern Discovery

The pattern discovery step consists of using algorithms which automatically discover patterns from the pre-processed data and the elicited domain knowledge. The choice of algorithm depends on the knowledge discovery goal. Due to the large amounts of Internet data from which knowledge is to be discovered, the algorithms used in this stage must be efficient. It is usually technically infeasible that this task is totally automated and independent of user intervention. Different paradigms require different parameters to be set by the user. Example parameters are number of hidden layers, number of nodes per layer and various learning parameters like learning rate and error tolerance for neural networks, population size, selection, mutation and cross-over probabilities for genetic algorithms, membership functions in fuzzy systems, support and

confidence thresholds in association algorithms and so on. Tuning these parameters is normally an iterative process and forms part of the refinement step (see Section 2.9).

In the context of electronic commerce, effectiveness measures have often to be redefined. In addition to standard measures such as support and confidence, the modification usually incorporates marketing-specific measures, for example, the lift ratio which is defined as $\text{lift} = \frac{\text{P(class} \mid \text{sample)}}{\text{P(class} \mid \text{population)}}$. Other measures which have been applied are cumulative lift and lifetime values (Berry & Linoff, 1997).

2.8 Knowledge Post-processing

Trivial and obsolete information must be filtered out, which is usually performed by ordering the knowledge and threshold based on some ranking. The ranking is based on support, confidence and interestingness measures of the knowledge. Additionally, in direct marketing, the values are applied on deciles, which allows decile analysis.

Discovered knowledge must be presented in a user-readable way, using either visualisation techniques or natural language constructs. One typical electronic commerce scenario is to provide a marketing expert with patterns s/he can interpret and verify, which have to be in the form of rules or visuals; another is to use the output as input for online prediction and dynamic creation of web pages, depending on the outcome of the built model.

A further aspect of knowledge post-processing is knowledge validation. Knowledge must be validated before it can be used for marketing actions on the Internet, for example, roll-outs. The most commonly used techniques here are holdout sampling, random resampling, *n*-fold cross-validation, and bootstrapping (Anand & Büchner, 1998).

Due to the fact that Internet data used as input to the knowledge discovery process is dynamic and prone to updates, the discovered patterns have to be maintained. Setting up a knowledge maintenance mechanism consists of re-applying the already set up process for the particular problem or using an incremental methodology that updates the knowledge as the data logged changes.

2.9 Refinement Process

The examination of the knowledge by marketing experts may lead to the refinement process, during which the domain knowledge as well as the actual goal posts of the discovery may be modified. Refinement can take the form of redefining the Internet and marketing data used in the discovery, a change in the methodology used, the user defining additional constraints on the algorithm(s), modification of the marketing knowledge used or calibration of parameters. Once the refinement is completed, the pattern discovery and knowledge post-processing stages are repeated. Note that the refinement process is not a stage of the data mining process. Instead it constitutes its iterative aspects and may make use of the initial stages of the process, that is data prospecting, methodology identification, domain knowledge elicitation, and data pre-processing.

3. Applications of the Internet-enabled Knowledge Discovery Process

Marketing experts divide the customer relationship life-cycle into four distinct steps, which cover attraction, retention, cross-sales, and departure. It has been recognised that mass marketing techniques are, exceptions excluded, inappropriate for e-commerce scenarios. More successful are direct marketing strategies, supported by knowledge discovery techniques (Ling & Li, 1998). A knowledge discovery scenario is presented for all four periods, each of which covers the discovery goal, marketing strategy, and data mining approach[2], which have been performed through the application of the outlined Internet-related process.

3.1 Customer Attraction

The two essential parts of attraction are the selection of new prospective customers and the acquisition of selected potential candidates. One marketing strategy to perform this exercise, among others, is to find common characteristics in already existing visitors' information and behaviour for the classes of profitable and non-profitable customers. These groups are then used as labels for a classifier to discover Internet marketing rules, which are applied online on site visitors. Depending on the outcome, a dynamically created page is displayed, whose contents depends on found associations between browser information and offered products / services.

The three classification labels used were 'no customer', that is browsers who have logged in, but did not purchase, 'visitor once' and 'visitor regular'. An example rule is as follows.

```
if Region = IRL and
   Domain1 IN [uk, ie] and
   Session > 320 Seconds
then VisitorRegular
Support = 6,4%; Confidence = 37,2%
```

This type of rule can then be used for further marketing actions such as displaying special offers to first time browsers from the two mentioned domains after they have spent a certain period of time on the shopping site.

3.2 Customer Retention

Customer retention is the step of managing the process of keeping the online shopper as loyal as possible. Due to the non-existence of physical distances between providers, this is an extremely challenging task in electronic commerce scenarios. One strategy is similar to that of acquisition, that is dynamically creating web offers based on associations. However, it has been proven more successful to consider associations across time, also known as sequential patterns. Typical

---

[2] No methodology-related information (see Section 2.5) is provided, since this choice is of generic nature and neither Internet- nor electronic commerce-specific.

sequences in electronic commerce data are representing navigational behaviour of shoppers in the forms of page visit series (Chen *et al*., 1996).

Agrawal & Srikant (1995)'s a priori algorithm has been extended so it can handle duplicates in sequences, which is relevant to discover navigational behaviour. The M*i*DAS[3] algorithm (Büchner *et al*., 1999b) also supports domain knowledge in the form of multi-level concept hierarchies, networks, and page chains (see Section 2.4). A found sequence looks as following.

```
{
ecom.infm.ulst.ac.uk/,
ecom.infm.ulst.ac.uk/News_Resources.html,
ecom.infm.ulst.ac.uk/Journals.html,
ecom.infm.ulst.ac.uk/,
ecom.infm.ulst.ac.uk/search.htm,
}
Support = 3.8%; Confidence = 31.0%
```

The discovered sequence can then be used to display special offers dynamically to keep a customer interested in the site, after a certain page sequence with a threshold support and / or confidence value has been visited.

3.3 Cross-Sales

The objective of cross-sales is to diversify selling activities horizontally and / or vertically to an existing customer base. We have adopted a traditional generic cross-sales methodology (Anand *et al*., 1998), in order to perform the given task in an electronic commerce environment.

For discovering potential customers, characteristic rules of existing cross-sellers had to be discovered, which was performed through the application of attribute-orientated induction. For a scenario in which the product CD is being cross-sold to book sellers, an example rule is

```
if Product = book then
   Domain1 = uk and
   Domain2 = ac and
   Category = Tools
Support = 16.4%; Interest = 0.34
```

Deviation detection is used to calculate the interest measure and to filter out the less interesting rules. The entire set of discovered interesting rules can then be used as the model to be applied at run-time on incoming actions and requests from existing customers.

---

[3] M*i*DAS stands for Mining Internet Data for Associative Sequences.

3.4 Customer Departure

Customers who depart have either stopped purchasing a certain service or product and / or have moved to a competitor, which is also known as churn. The goal of customer departure prediction is to take action in order to prevent the exit (for instance, through a targeted promotion) or to prevent further costs in case the customer will leave, no matter what action will be taken.

Since a customer in an electronic commerce scenario does not explicitly leave, a user-defined delta value has to chosen as a threshold in which no activities have been recorded (neither browsing nor purchases). Log files from a certain period previous to the last activity have then to be analysed similarly to the customer retention scenario, that is sequences are discovered in order to find characteristics of churners. In parallel, classification exercises can be performed on the customer data in order to distinguish leavers from current customers. The types of patterns discovered are similar to the ones shown in sections 3.1 to 3.3 and are omitted for reasons of brevity.

4. Related Work

Etzioni (1996) has suggested three types of web mining activities, viz. *resource discovery*, usually carried out by intelligent agents, *information extraction* from newly discovered pages, and *generalisation*. For the purpose of the discussion of related work only the latter category is considered, since it has the most important impact on electronic commerce research.

Zaïane *et al*. (1998) have applied various traditional OLAP and data mining techniques to Internet log files in order to find different types of patterns, which can be harnessed as electronic commerce decision support knowledge. The process involves a data cleansing and filtering stage (manipulation of date and time related fields, removal of futile entries, et cetera), which is followed by a transformation step that reorganises log entries supported by meta data. The pre-processed data is then loaded into a data warehouse which has an *n*-dimensional web log cube as basis. From this cube, various standard OLAP techniques are applied, such as drill-down, roll-up, slicing, and dicing. Additionally, artificial intelligence and statistically-based data mining techniques are applied on the collected data which include characterisation, discrimination, association, regression, classification, and sequential patterns.

Cooley *et al*. (1997) have developed a similar, but more powerful process. It includes intelligent cleansing (outlier elimination and removal of irrelevant values) and pre-processing (user and session identification, path completion, reverse DNS lookups, et cetera) of Internet log files, as well as the creation of data warehousing-like views (Cooley *et al*., 1999). In addition to Zaïane *et al*. (1998)'s approach, registration data, as well as transaction information is integrated in the materialised view. From this view, various data mining techniques can be applied; named are path analysis, associations, sequences, clustering and classification. These patterns can then be analysed using OLAP tools, visualisation mechanisms or knowledge engineering techniques.

Both approaches share some obstacles which limit their applicability in several ways. Firstly, they only support a sub-set of Internet data sources, which has proven insufficient for real-world electronic commerce exploitation. Secondly, no domain knowledge (marketing expertise) has

been incorporated in either web mining processes, which we see as an essential feature. Thirdly, the discovery of web access patterns is designed as a one-way flow, rather than a feedback process, that is the discovered knowledge does not feed back in the knowledge discovery process. And lastly, both approaches are very data mining-biased, in that they re-use existing techniques which have not been tailored towards Internet and electronic commerce purposes.

5. Conclusions and Further Work

We have proposed a holistic approach to discover marketing intelligence from Internet data in the shape of a process. The electronic commerce-enabled knowledge discovery process contains the steps human resource identification, problem specification, data prospecting, domain knowledge elicitation, methodology identification, data pre-processing, pattern discovery and knowledge post-processing. It also involves the three types of expertise required in during a project, namely a web administrator, a marketing expert, and a data mining expert. To show the validity and applicability of the proposed approach, the electronic commerce marketing scenarios customer attraction, customer retention and cross-sales have been tackled with the outlined process.

Current work involves the incorporation of more sophisticated domain knowledge and syntactic constraints (such as Büchner *et al*., 1999b), better transaction support, similar to Cooley *et al*. (1999), and the scalability improvement of developed web-enabled data mining algorithms. The work proposed in here exclusively focuses on customer-to-business relationships on the Internet. A (financially) more attractive, but currently less mature discipline, is business-to-business electronic commerce. The outlined Internet-enabled process is currently extended for handling modelling activities in pure business-orientated scenarios.

6. References

[AB98] S.S. Anand, A.G. Büchner. Decision Support through Data Mining, FT Pitman Publishers, 1998.

[ABH95] S.S. Anand, D.A. Bell, J.G. Hughes. The Role of Domain Knowledge in Data Mining, *Proc. 4th Int'l. ACM Conf. on Information and Knowledge Management*, pg. 37-43, 1995.

[APHB98] S.S. Anand, A.R. Patrick, J.G. Hughes, D.A. Bell. A Data Mining Methodology for Cross Sales, Knowledge-based Systems Journal, 10:449-461, 1998.

[AS95] R. Agrawal, R. Srikant. Mining Sequential Patterns, *Proc. 11th Int'l Conf. on Data Engineering*, pp. 3-14, 1995.

[BBM+99] A.G. Büchner, M. Baumgarten, M.D. Mulvenna, S.S. Anand, J.G. Hughes, Discovering Marketing-driven Navigation Patterns, submitted for publication, 1999.

[BBH99] A.G. Büchner, J.G. Hughes, D.A. Bell. Contextual Domain Knowledge for Incorporation in Data Mining Systems, *Proc. AAAI Workshop on Reasoning in Context for AI Applications*, 1999.

[Bha98] S. Bhattacharyya. Direct Marketing Response Models using Genetic Algorithms, *Proc. 4th Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 144-148, 1998.

[Big96] J.P. Bigus. Data Mining with Neural Networks, McGraw Hill, 1996.

[BM98] A.G. Büchner, M.D. Mulvenna. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining, *ACM SIGMOD Record*, 27(4):54-61, 1998.

[BL97] M.J.A. Berry, G. Linoff. Data Mining Techniques for Marketing, Sales, and Customer Support, Wiley Computer Publishing, 1997.

[BS97] A. Berson, S.J. Smith. Data Warehousing, Data Mining and OLAP, McGraw Hill, 1997.

[CD97] S. Chaudhuri, U. Dayal. An Overview of Data Warehousing and OLAP Technology, Technical Report MSR-TR-97-14, Microsoft Research, 1997.

[CMS97] R. Cooley, B. Mobasher, J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web, *Proc. 9th IEEE Int'l Conf. on Tools with Artificial Intelligence*, 1997.

[CMS99] R. Cooley, B. Mobasher, J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns, *Knowledge and Information Systems*, 1(1), 1999.

[CPY96] M.S. Chen, J.S. Park, P.S. Yu. Data Mining for Traversal Patterns in a Web Environment, *Proc. 16th Intl'l Conf. on Distributed Computing Systems,* pp. 385-392, 1996.

[Etz96] O. Etzioni. The World-Wide Web: Quagmire or Gold Mine?, *Comm. of the ACM*, 39(11):65-68, 1996.

[HRU96] V. Harinarayan, A. Rajarman, J.D. Ullman. Implementing data cubes efficiently, *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pp. 205-216, 1996.

[LL98] C.X. Ling, C. Li. Data Mining for Direct Marketing: Problems and Solutions, *Proc. 4th Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 73-79, 1998.

[MNB98] M.D. Mulvenna, M.T. Norwood, A.G. Büchner. Data-driven Marketing, *Int'l Journal of Electronic Marketing*, 8(3):32-35, 1998.

[ZXH98] O. R. Zaïane, M. Xin, J. Han. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs, *Proc. Advances in Digital Libraries Conf.*, pp. 19-29, 1998.